This manuscript is currently in press at the *Journal of Experimental Psychology: Human Perception and Performance*. Data and analysis code are available on the OSF (https://osf.io/jurcq/).

Comparing theory-driven and data-driven attractiveness models using images of real women's faces

Iris J Holzleitner[1], Anthony J Lee[2], Amanda C Hahn[3], Michal Kandrik[4], Jeanne Bovet[5], Julien P Renoult[6], David Simmons[7], Oliver Garrod[1], Lisa M DeBruine[1], Benedict C Jones[1]

[1] Institute of Neuroscience and Psychology, University of Glasgow, UK

[2] Division of Psychology, University of Stirling, UK

[3] Department of Psychology, Humboldt State University, USA

[4] Department of Experimental and Applied Psychology, Vrije Universiteit Amsterdam, Netherlands

[5] Stony Brook University, NY, USA

[6] Centre of Evolutionary and Functional Ecology (CEFE UMR5175, CNRS—University of Montpellier—University Paul-Valery Montpellier—EPHE), France

[7] School of Psychology, University of Glasgow, UK

Author Note

This research was supported by European Research Council grants awarded to BCJ (OCMATE), LMD (KINSHIP) and AJL (MULTIPREF). We wish to thank Alexander Todorov, Alex Jones and two anonymous reviewers for their helpful comments on an earlier draft of this manuscript.

Author Contributions

IJH, AJL, BCJ, LMD and DS designed the study. ACH and MK collected the data. IJH conducted shape analysis, AJL conducted colour analysis, and JB, JPR and OG conducted sparseness analysis. IJH, AJL, LMD and BCJ carried out general data analysis and drafted the manuscript, which was revised by all authors.

Word count: 2,979

Abstract

Facial attractiveness plays a critical role in social interaction, influencing many different social outcomes. However, the factors that influence facial attractiveness judgments remain relatively poorly understood. Here, we used a sample of 594 young adult female face images to compare the performance of existing theory-driven models of facial attractiveness and a data-driven (i.e., theory-neutral) model. Our data-driven model and a theory-driven model including various traits commonly studied in facial attractiveness research (asymmetry, averageness, sexual dimorphism, body mass index, and representational sparseness) performed similarly well. By contrast, univariate theory-driven models performed relatively poorly. These results (1) highlight the utility of data driven models of facial attractiveness and (2) suggest that theory-driven research on facial attractiveness would benefit from greater adoption of multivariate approaches, rather than the univariate approaches that they currently almost exclusively employ.

Keywords: mate preferences, principal component analysis, face perception, face processing

Comparing theory-driven and data-driven attractiveness models using images of real women's faces

Faces are a particularly important feature for social communication (Haxby et al., 2000; Little et al., 2011; Todorov et al., 2015) and facial attractiveness influences important social outcomes (Langlois et al., 2000; Rhodes, 2006). For example, people generally prefer to form romantic and platonic relationships with facially attractive individuals and even prefer to hire and vote for individuals with attractive faces (Langlois et al., 2000; Rhodes, 2006). Thus, understanding the factors that determine facial attractiveness can provide insights into an attribute that appears to have critical effects on social interactions and outcomes (Little et al., 2011; Rhodes, 2006).

Studies of the facial characteristics that influence attractiveness judgments have typically employed a top-down (i.e. theory-driven) approach in which possible relationships between attractiveness ratings of faces and specific facial characteristics, such as asymmetry, averageness or sexual dimorphism, are tested (Little et al., 2011; Rhodes, 2006). However, when objectively assessed from face images, these facial characteristics reliably explain only a small proportion of the variance in attractiveness ratings (Said & Todorov, 2011).

By contrast with the top-down approach described above, Said and Todorov (2011) used a bottom-up approach to study the characteristics that influence facial attractiveness judgments. Said and Todorov (2011) predicted attractiveness from the position a face occupied in face space. Face space is a multi-dimensional space representing global shape and color properties of

faces derived from Principal Component Analysis (see, e.g., O'Toole et al., 2018 for a recent review). Crucially, Said and Todorov's (2011) 'face space' model had considerably greater predictive power than a top-down model including averageness and sexual dimorphism. For female faces, the R^2 for the top-down model was .37 and for the face space model .62. For male faces, the R^2 for the top-down model was .04 and for the face space model .71. However, there are two important limitations to Said and Todorov's (2011) study.

First, as Said and Todorov (2011) emphasized in their discussion, their study used attractiveness ratings of synthetic faces. Such ratings may not be representative of how people rate the attractiveness of photographs of real faces. Indeed, some research suggests that social judgments of synthetic faces can be qualitatively different from social judgments of photographs of real faces (Balas & Pacella, 2017; Balas, Tupa & Pacella, 2018). Synthetic faces are also processed differently (as evidenced by differences in recognition rates) from photographs of real faces (Balas & Pacella, 2015; Kätsyri, 2018). Thus, replicating Said and Todorov's (2011) bottom-up approach using photographs of real faces is essential to establish whether their findings for synthetic faces generalize to photographs of real faces.

Second, since Said and Todorov (2011) conducted their study, new research has identified further specific facial characteristics that are claimed to be good predictors of facial attractiveness. For example, several studies have reported that body mass index (BMI) is an important predictor of women's facial attractiveness, potentially because it is an important health cue (Coetzee et al., 2009; Han et al., 2016; Rantala et al., 2013). Other work

reported that measures of representational sparseness are good predictors of women's facial attractiveness. These measures are derived from algorithms that estimate the sparseness of neurons in the visual cortex required to represent a given face image and are thought to predict attractiveness because they index image-coding efficiency (Renoult et al., 2016). In other words, the sparseness of the activity of simple cells in V1 can be estimated for individual face images and is positively correlated with attractiveness (Renoult et al., 2016).

Thus, comparing Said and Todorov's (2011) face-space model with more recent top-down models including facial characteristics absent from the original study (BMI and representational sparseness) is essential to establish the superiority of the bottom-up approach over top-down models.

In light of the above, we compared the performance of top-down models of facial attractiveness that included measured asymmetry, averageness, sexual dimorphism, BMI, and representational sparseness as predictors, to a bottom-up model using shape and color principal components derived from a principal component analysis of our face stimuli. Rather than using synthetic face images, we analyzed face photographs of 594 young adult women.

Methods

Face images

We recruited 594 young adult women for the study (mean age=21.5 years, SD=3.2 years). All participants were students at the University of Glasgow, participating as part of a larger project on hormones and mating psychology (Jones et al., 2018a; Jones et al., 2018b; Jones et al., 2018c). Each woman

first cleaned her face with hypoallergenic face wipes to remove any make up. A full-face digital photograph was taken a minimum of 10 minutes later. Photographs were taken in a small windowless room against a constant background, under standardized diffuse lighting conditions, and participants were instructed to pose with a neutral expression. Camera-to-head distance and camera settings were held constant. Participants wore a white smock covering their clothing when photographed to control for possible effects of reflectance from clothing. Photographs were taken using a Nikon D300S digital camera with an AF Micro-Nikkor 60mm f/2.8D lens. A GretagMacbeth 24-square ColorChecker chart was included in each image for use in color calibration.

Following Jones et al. (2015), face images were color-calibrated using a least-squares transform from an 11-expression polynomial expansion developed to standardize color information across images (Hong et al., 2001). Each image was masked so that hairstyle and clothing were not visible and placed on a white background.

Facial attractiveness ratings

The 594 face images were then rated for attractiveness using a 1 (much less attractive than average) to 7 (much more attractive than average) scale by 16 men and 16 women. Trial order was fully randomized. Inter-rater agreement was high for these ratings (ICC=0.30, 95% CI [0.21, 0.43]; Cronbach's alpha=.93, 95% CI [.92, .94]) and ratings by male and female raters were highly correlated (r=.87, 95% CI [0.85, 0.89], N=594, p<.001). Consequently we calculated an average attractiveness score for each image after standardizing ratings for each rater (M=−1.57, SD=0.57) to use in our

analyses. Ratings were standardized prior to averaging to account for individual differences in scale use and because this was done in Said and Todorov's original study. Images were standardized on pupil positions prior to rating.

The numbers of raters was chosen based on simulations (see https://osf.io/x7fus/) sampling from a population of 2513 raters, each of whom had rated the attractiveness of 102 faces. More than 99% of 1000 random samples of 15 raters produced Cronbach's alphas >.8, indicating high reliability of ratings at the mean rating level (90% of all alphas were >.85). Furthermore, increasing the number of raters providing attractiveness ratings has a negligible effect on the mean attractiveness ratings once ratings have been collected from 28 raters (Hehman et al., 2018).

Principal Component Analysis (PCA) of face images

Shape principal components (PCs) were derived from 132 Procrustes-aligned points on each of the 594 faces using a method described in Wolffhechel et al. (2015). Color PCs were derived from the RGB values for each pixel from shape-normalized images. Non-face regions of the images were masked prior to the PCA. To avoid overfitting, we used the broken stick criterion to select PCs to be included as predictors in our analyses (see Jackson, 1993, for a discussion of the advantages of this criterion). The broken stick method partitions the total variance ("the stick") into as many segments as there are PCs, assigning each segment a proportionally increasing amount of variance. PCs are retained if their eigenvalue is greater than that of the corresponding

segment from the broken stick model. Code for deriving and extracting principal components is available at https://osf.io/jurcq/.

Measuring averageness

Facial averageness was measured from each photograph using a technique adapted from Lee et al. (2016), separately for shape and color. This method uses the shape/color components to measure the distance the face lies from the mathematical average shape/color for the sample of faces. That is, the average shape/color values for the sample are calculated and, for each image, the Euclidean distance to the average is derived. Higher scores indicate more distinctive face shapes. Code for calculating distinctiveness scores is available at https://osf.io/jurcq/.

Measuring sexual dimorphism

Facial sexual dimorphism was measured objectively from each photograph, separately for shape and color. Sexual dimorphism scores were calculated using a vector analysis method (e.g., Komori et al., 2011). This method uses shape/color principal components to locate each face on a female-male continuum. The female-male continuum was defined by calculating the average shape/color information of 50 male (mean age=20.85 years, SD=3.01 years) and the average of 50 female (mean age=20.60 years, SD=1.38 years) faces. These faces were not included in the main sample. Sexual dimorphism scores were then derived by projecting each image onto this female-male vector. Code for calculating sexual dimorphism scores is available at https://osf.io/jurcq/.

Measuring asymmetry

Facial asymmetry was measured from each photograph using a technique adapted from Komori et al. (2009). Facial asymmetry was measured in shape only. For each image, the landmark template was mirrored, and asymmetry measured as the Euclidean distance between original and mirrored templates. Code for calculating asymmetry is available at https://osf.io/jurcq/.

Measuring sparseness

Following Renoult et al. (2016), we used an algorithm that estimates the sparseness of neurons in the visual cortex responses that would be needed to represent images of female faces. This algorithm uses a feature dictionary based on Olshausen and Field's (1997) work on the properties of receptive fields in the visual cortex. Also following Renoult et al. (2016), we defined sparseness of the encoding as the kurtosis of the estimated feature coefficients. Our MATLAB code for calculating sparseness is publicly available at https://osf.io/jurcq/.

Measuring Body Mass Index (BMI)

Height (M=165.8 cm, SD=6.3 cm) and weight (M=63.8 kg, SD=11.9 kg) were measured for the women who had been photographed. These measurements were used to calculate BMI (M=23.1 kg/m^2, SD=3.8 kg/m^2). Forty-nine women chose not to provide both measurements, so BMI could not be calculated for these women.

Results

Analyses were conducted using *R v3.5.1* (R Core Team, 2018) and the packages *geomorph v3.0.6* (Adams et al., 2018) for morphometric analyses and *caret v6.0-79* (Kuhn, 2008) for cross-validation analyses. All data and analysis code are publicly available at https://osf.io/jurcq/.

First, we specified seven models, each with attractiveness rating as the dependent variable. The *asymmetry* model had linear shape asymmetry scores as the predictor. The *averageness* model had linear color and shape averageness scores as the predictors. The *sexual dimorphism* model had both linear and quadratic color and shape sexual dimorphism scores as predictors. The quadratic terms were included because of previous research suggesting a quadratic relationship between sexual dimorphism and attractiveness (DeBruine et al., 2007; Holzleitner & Perrett, 2017). The *BMI* model had both linear and quadratic BMI as the predictors. The quadratic term was included because of previous research suggesting a quadratic relationship between BMI and attractiveness (Coetzee et al., 2009; Han et al., 2016; Rantala et al., 2013). The *sparseness* model had both linear and quadratic sparseness scores as the predictors. The quadratic term was included because exploratory analyses suggested there was a weak quadratic relationship between sparseness and attractiveness. The *top-down combined* model included all predictors from the averageness, sexual dimorphism, BMI, and sparseness models. Finally (and following Said & Todorov, 2011), the *bottom-up face space* model included linear and quadratic effects for all 12 shape and 60 color principal components that were selected for analyses

using the broken-stick criterion. Full model specifications are given in our supplemental materials and at https://osf.io/jurcq/.

Next, we used 10-fold cross validation with 10 repeats (i.e., 100 resamples) to estimate the root-mean-square error (RMSE) for each model (Figure 1). RMSE is the square root of the mean squared differences between predicted and observed values, and as such a measure of predictive accuracy. A value of 0 would indicate a perfect fit of the data. By contrast with R^2, RMSE is not inflated by the number of predictors. The RMSE for the top-down combined model was 0.47 (SD=0.04). The RMSE for the face space model was 0.46 (SD=.04). To test how much variance was uniquely explained by the top-down combined compared to the face space model, we re-tested top-down combined and face space models based on all 594 observations (instead of using cross-validation). Together, top-down combined and face space predictors explained 62.7% of the variance in attractiveness ratings. The top-down model explained 34.4% of the variance, with 7.2% variance being unique variance that was not explained by the face space model. The face space model explained 55.6% of the variance in attractiveness, with 28.4% variance being unique variance that was not explained by the top-down combined model (see supplemental materials).

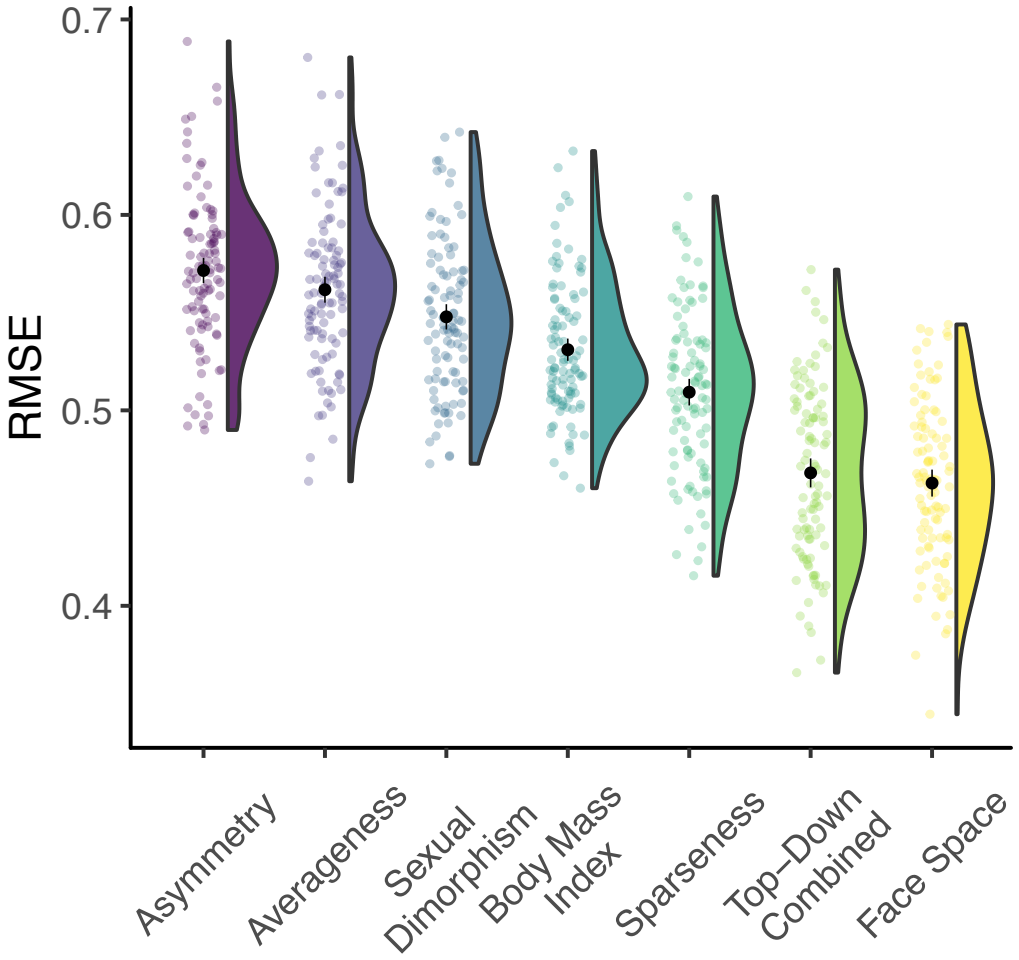

Figure 1. Raincloud plots showing performance of the seven models in predicting facial attractiveness. Black dots show the mean RMSE from 10-fold cross validation with 10 repeats. Black bars show 90% confidence intervals.

Additional models comparing all possible combinations of asymmetry, averageness, sexual dimorphism, BMI, and sparseness with the face space model are reported in our supplemental materials.

To test whether the face space model is over-fitting, we used the Akaike Information Criterion (AIC) as a measure of model fit and compared the AIC from each of the 100 resamples to those of the top-down combined model. Despite its higher number of predictors, the face space model showed a

better fit (average AIC=−770.99) than the top-down combined model (average AIC=−686.54) for each of the 100 resamples. The minimum difference in AIC was 39.10.

Finally, we used variable selection based on the AIC to identify PCs that were selected in all of the resamples. In other words, we used the AIC in a stepwise backward regression to determine the best predictors for each resample. We then visualized predictors that were retained in all 100 resamples. Fourteen PCs (four shape, 10 color) satisfied this criterion. The four shape PCs are visualized in Figure 2. The 10 color PCs are visualized in Figure 3.

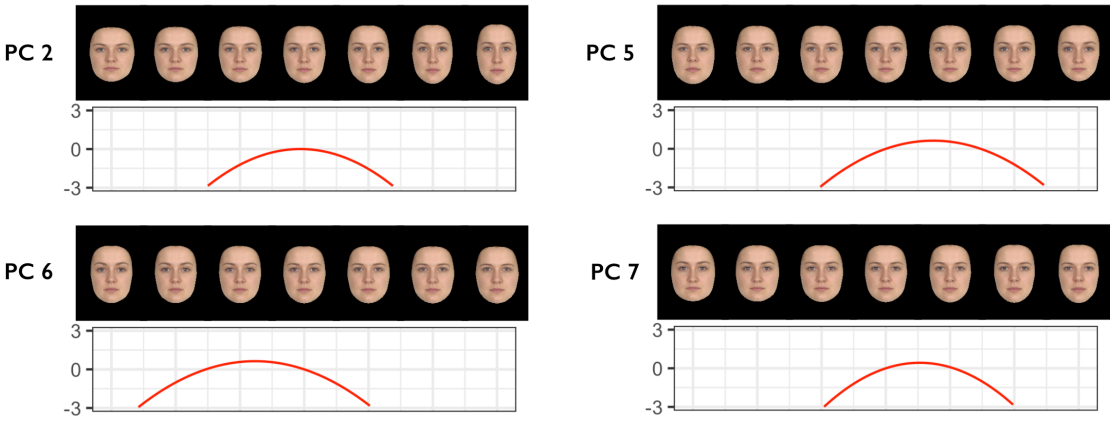

Figure 2. Visualization of the four shape PCs selected in all 100 resamples. The top row of each panel shows the variance in the respective PC ranging from -3 to +3 standard deviations. The bottom row in each panel shows the average relationship between the PC (x-axis) and standardized attractiveness ratings as predicted from the model (y-axis). The quadratic effects are graphed where they were significant in more than 80 resamples.

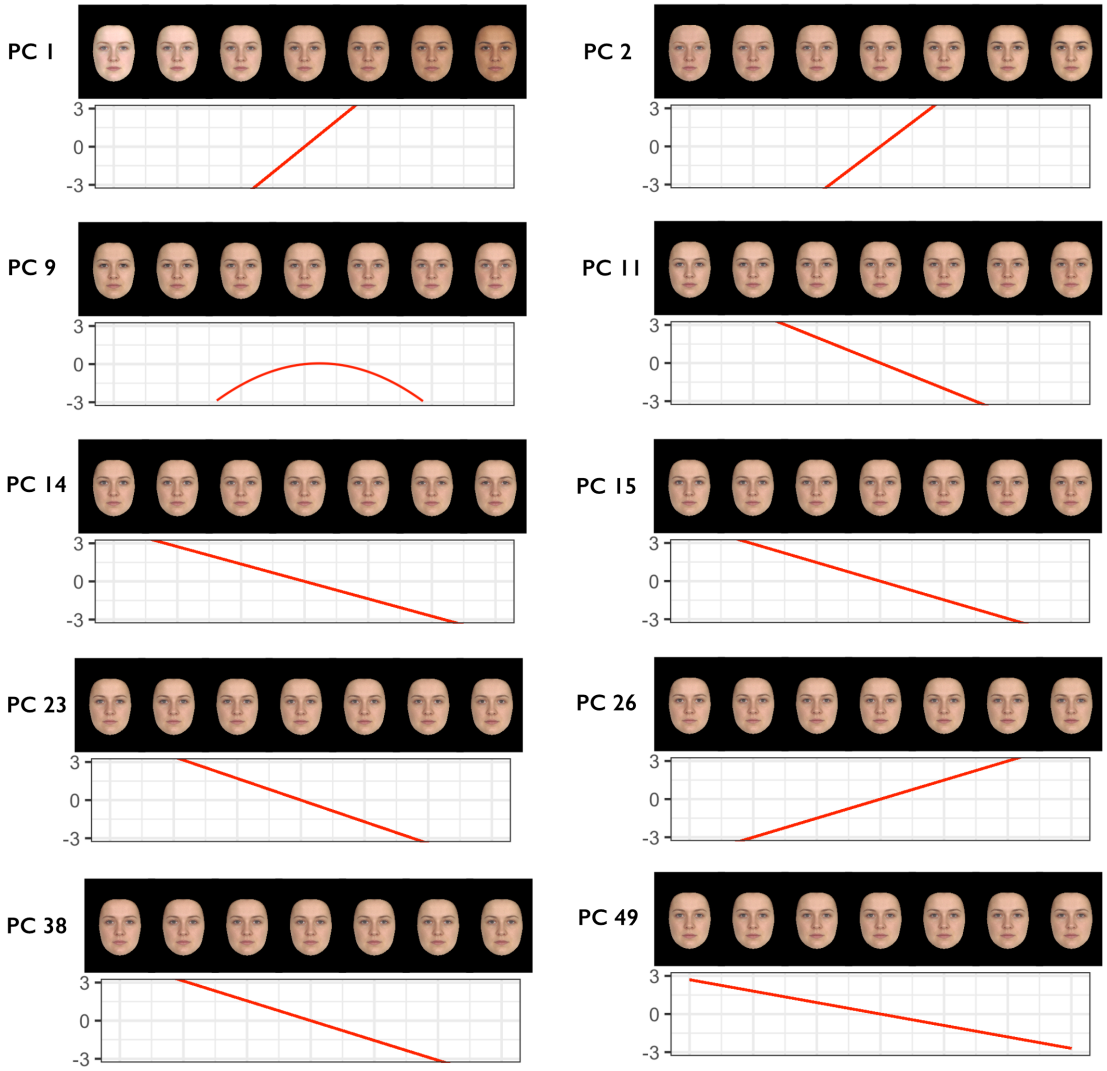

Figure 3. Visualization of the 10 color PCs selected in all 100 resamples. The top row of each panel shows the variance in the respective PC ranging from -3 to +3 standard deviations. The bottom row in each panel shows the average relationship between the PC (x-axis) and standardized attractiveness ratings as predicted from the model (y-axis). The quadratic effects are graphed where they were significant in more than 80 resamples.

Discussion

Based on the RMSE values, image sparseness was the best predictor of attractiveness among the top-down predictors in our study, then BMI, then sexual dimorphism, then averageness, then asymmetry (see Figure 1). This pattern of results is consistent with other recent work suggesting averageness

and sexual dimorphism are relatively unimportant for women's facial attractiveness (Said & Todorov, 2011), while BMI (Coetzee et al., 2009; Han et al., 2016; Rantala et al., 2013) and sparseness (Renoult et al., 2016) are relatively good predictors. Our study is the first to describe the effectiveness of these five different top-down predictors in a single sample of face images. Importantly, the large sample of images tested and the cross-validation methods we used for our analyses mean that our estimates of the predictive power of these characteristics are likely to be reliable and robust.

Our face-space model performed similarly to the combined top-down model (see RMSE in Figure 1). Only the performance of the combined top-down model came close to the performance of the face-space models. These results demonstrate the utility of face-space models for studying facial attractiveness and highlight the limitations of individual (i.e., univariate) theory-driven models. These results are consistent with other recent work finding no evidence that hormone levels or susceptibility to illnesses, the underlying characteristics that these top-down predictors are assumed to signal, are correlated with facial attractiveness in young adult women (Cai et al., 2018; Jones et al., 2018d). Importantly, the AIC showed that the face-space model was a better fit than the combined top-down model, indicating that the face-space and combined top-down models' comparable performance was not a consequence of overfitting in the face space model. Note that both the face-space and combined top-down models included both shape and color predictors.

Visualizations of the PCs revealed shape and color components of attractiveness that are not typically emphasized in research on facial

attractiveness. For example, face elongation (shape PC2) and the ratio of feature size to face size (shape PC5) appear to be important predictors of attractive face shapes (see Figure 2). For color, skin tone (color PC1) and feature contrast (color PC2) appear to have strong effects on women's facial attractiveness (see Figure 3). These patterns complement Said and Todorov's (2011) results for synthetic faces and also previous work highlighting the importance of color information for facial attractiveness (e.g., Russell et al., 2016; Stephen et al., 2009; but see Foo et al., 2017). For example, the color PCs that predict attractiveness best may relate to cues considered in theory-driven studies of attractiveness, such as carotenoid-related skin tone and sexually dimorphic contrast information (Henderson et al., 2018; Jones, 2018; Russell et al., 2016; Stephen et al., 2009). Shape PCs that predict attractiveness best (e.g., face elongation) may be those related to height (Re et al., 2013, Mitteroecker et al., 2013). The curvilinear relationship observed for many components of the face space model may be consistent with claims that facial attractiveness is influenced, at least in part, by aversions to specific extreme facial characteristics (e.g., Zebrowitz & Rhodes, 2004). Whether these PCs reflect other theory-derived attractive facial cues that were not considered in our study (e.g., cues of residual fertility, Bovet et al., 2018) is an open empirical question. Whether our results generalize beyond the type of sample we tested here (young, predominantly white, female faces) is also an open empirical question.

Like Said and Todorov's (2011) pioneering work on statistical models of facial attractiveness, our results highlight how poorly many existing top-down models of facial attractiveness perform, at least when these top-down models

are univariate. This is the case even for female facial attractiveness, for which top-down models have been hypothesized to be particularly useful (Rhodes, 2006; Thornhill & Gangestad, 1999; Little et al., 2011) and is not simply an artefact of using the type of synthetic face images employed in Said and Todorov's (2011) original study.

In conclusion, we show that a model combining *multiple* theory-derived predictors can perform as well as a data-driven, face space model. However, this approach (combining multiple theory-derived predictors) is very uncommon in the facial attractiveness literature. We strongly suggest that research using theory-driven models to study facial attractiveness would benefit from this type of multivariate approach, rather than the univariate approach that they almost exclusively employ.

References

Adams, D. C., Collyer, M. L., & Kaliontzopoulou, A. (2018). Geomorph: Software for geometric morphometric analyses (Version 3.0.6). Retrieved from https://cran.r-project.org/package=geomorph

Balas, B., Tupa, L., & Pacella, J. (2018). Measuring social variables in real and artificial faces. *Computers in Human Behavior, 88*, 236-243. https://doi.org/10.1016/j.chb.2018.07.013

Balas, B., & Pacella, J. (2015). Artificial faces are harder to remember. *Computers in Human Behavior, 52*, 331-337. https://doi.org/10.1016/j.chb.2015.06.018

Balas, B., & Pacella, J. (2017). Trustworthiness perception is disrupted in artificial faces. *Computers in Human Behavior, 77*, 240-248. https://doi.org/10.1016/j.chb.2017.08.045

Bovet, J. , Barkat-Defradas, M. , Durand, V. , Faurie, C. & Raymond, M. (2018). Women's attractiveness is linked to expected age at menopause. *Journal of Evolutionary Biology, 31*, 229-238. https://doi.org/10.1111/jeb.13214

Cai, Z., Hahn, A. C., Zhang, W., Holzleitner, I. J., Lee, A. J., DeBruine, L. M., & Jones, B. C. (2018). No evidence that facial attractiveness, femininity, averageness, or coloration are cues to susceptibility to infectious illnesses in a university sample of young adult women. *Evolution and Human Behavior*. https://doi.org/10.1016/j.evolhumbehav.2018.10.002

Coetzee, V., Perrett, D. I., & Stephen, I. D. (2009). Facial adiposity: A cue to health? *Perception, 38*, 1700-1711. https://doi.org/10.1068/p6423

DeBruine, L. M., Jones, B. C., Unger, L., Little, A. C., & Feinberg, D. R. (2007). Dissociating averageness and attractiveness: Attractive faces are not always average. *Journal of Experimental Psychology: Human Perception and Performance, 33*(6), 1420-1430. https://doi.org/10.1037/0096-1523.33.6.1420

Foo, Y. Z., Simmons, L. W., & Rhodes, G. (2017). Predictors of facial attractiveness and health in humans. *Scientific Reports, 7*, 39731. https://doi.org/10.1038/srep39731

Han, C., Hahn, A. C., Fisher, C. I., DeBruine, L. M., & Jones, B. C. (2016). Women's facial attractiveness is related to their body mass index but not their salivary cortisol. *American Journal of Human Biology, 28*(3), 352-355. https://doi.org/10.1002/ajhb.22792

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences, 4*(6), 223-233. https://doi.org/10.1016/S1364-6613(00)01482-0

Hehman, E., Xie, S. Y., Ofosu, E. K., & Nespoli, G. A. (2018). Assessing the point at which averages are stable: A tool illustrated in the context of person perception. *PsyArXiv*. https://doi.org/10.17605/OSF.IO/2N6JQ

Henderson A. J., Holzleitner I. J., Talamas S. N., & Perrett D. I. (2016). Perception of health from facial cues. Philosophical Transactions of the Royal Society B, 371, 20150380. https://doi.org/10.1098/rstb.2015.0380

Holzleitner, I. J., & Perrett, D. I. (2017). Women's Preferences for Men's Facial Masculinity: Trade-Off Accounts Revisited. *Adaptive Human Behavior and Physiology, 3*(4), 304-320. https://doi.org/10.1007/s40750-017-0070-3

Hong, G., Luo, M. R., & Rhodes, P. A. (2001). A study of digital camera colourimetric characterization based on polynomial modeling. *Colour Research & Application, 26*(1), 76-84. https://doi.org/10.1002/1520-6378(200102)26:1<76::AID-COL8>3.0.CO;2-3

Jackson, D. A. (1993). Stopping Rules in Principal Components Analysis: A Comparison of Heuristical and Statistical Approaches. *Ecology, 74*(8), 2204-2214. https://doi.org/10.2307/1939574

Jones, A. L. (2018). The influence of shape and colour cue classes on facial health perception. *Evolution and Human Behavior, 39*(1), 19-29. https://doi.org/10.1016/j.evolhumbehav.2017.09.005

Jones, B. C., Hahn, A. C., Fisher, C. I., Wang, H., Kandrik, M., & DeBruine, L. M. (2018a). General sexual desire, but not desire for uncommitted sexual relationships, tracks changes in women's hormonal status. *Psychoneuroendocrinology, 88*, 153-157. https://doi.org/10.1016/j.psyneuen.2017.12.015

Jones, B. C., Hahn, A. C., Fisher, C. I., Wang, H., Kandrik, M., Han, C., . . . DeBruine, L. M. (2018b). No Compelling Evidence That Preferences for Facial Masculinity Track Changes in Women's Hormonal Status. *Psychological Science*, 0956797618760197. https://doi.org/10.1177/0956797618760197

Jones, B. C., Hahn, A. C., Fisher, C. I., Wang, H., Kandrik, M., Lee, A. J., . . . DeBruine, L. M. (2018c). Hormonal correlates of pathogen disgust: testing the compensatory prophylaxis hypothesis. *Evolution and Human Behavior, 39*(2), 166-169. https://doi.org/10.1016/j.evolhumbehav.2017.12.004

Jones, B. C., Hahn, A. C., Fisher, C. I., Wang, H., Kandrik, M., Lao, J., . . . DeBruine, L. M. (2018d). No compelling evidence that more physically attractive young adult women have higher estradiol or progesterone. *Psychoneuroendocrinology, 98*, 1-5. https://doi.org/10.1016/j.psyneuen.2018.07.026

Jones, B. C., Hahn, A. C., Fisher, C. I., Wincenciak, J., Kandrik, M., Roberts, S. C., . . . DeBruine, L. M. (2015). Facial colouration tracks changes in women's estradiol. *Psychoneuroendocrinology, 56*, 29-34. https://doi.org/10.1016/j.psyneuen.2015.02.021

Kätsyri, J. (2018). Those Virtual People all Look the Same to me: Computer-Rendered Faces Elicit a Higher False Alarm Rate Than Real Human Faces in a Recognition Memory Task. *Frontiers in Psychology, 9*, 1362. https://doi.org/10.3389/fpsyg.2018.01362

Komori M, Kawamura, S., & Ishihara, S. (2009). Averageness or symmetry: Which is more important for facial attractiveness? *Acta Psychologica, 131, 136-142.* https://doi.org/10.1016/j.actpsy.2009.03.008

Komori, M., Kawamura, S., & Ishihara, S. (2011). Multiple Mechanisms in the Perception of Face Gender: Effect of Sex-Irrelevant Features. *Journal of Experimental Psychology-Human Perception and Performance, 37*, 626-633. https://doi.org/10.1037/A0020369

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software, 28*(5). https://doi.org/10.18637/jss.v028.i05

Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hallam, M., & Smoot, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological Bulletin, 126*, 390-423. https://doi.org/10.1037/0033-2909.126.3.390

Lee, A. J., Mitchem, D. G., Wright, M. J., Martin, N. G., Keller, M. C., & Zietsch, B. P. (2016). Facial averageness and genetic quality: testing heritability, genetic correlation with attractiveness, and the paternal age effect. *Evolution and Human Behavior, 37*(1), 61-66. https://doi.org/10.1016/j.evolhumbehav.2015.08.003

Little, A. C., Jones, B. C., & DeBruine, L. M. (2011). Facial attractiveness: evolutionary based research. *Philosophical Transactions of the Royal Society B: Biological Sciences, 366*, 1638-1659. https://doi.org/10.1098/rstb.2010.0404

Mitteroecker, P., Gunz, P., Windhager, S., & Schaefer, K. (2013). A brief review of shape, form, and allometry in geometric morphometrics, with applications to human facial morphology. *Hystrix, the Italian Journal of Mammalogy, 24,* 59–66.

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research, 37*(23), 3311-3325. https://doi.org/10.1016/S0042-6989(97)00169-7

O'Toole, A. J., Castillo, C. D., Parde, C. J., Hill, M. Q., & Chellappa, R. (2018). Face Space Representations in Deep Convolutional Neural Networks.

Trends in Cognitive Sciences, 22(9), 794-809. https://doi.org/10.1016/j.tics.2018.06.006

R Core Team. (2018). R: A language and environment for statistical computing (Version 3.5.0). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rantala, M. J., Coetzee, V., Moore, F. R., Skrinda, I., Kecko, S., Krama, T., . . . Krams, I. (2013). Facial attractiveness is related to women's cortisol and body fat, but not with immune responsiveness. *Biology Letters, 9*(4). https://doi.org/10.1098/rsbl.2013.0255

Re, D. E., Hunter, D. W., Coetzee, V., Tiddeman, B. P., Xiao, D., DeBruine, L. M., Jones, B. C., & Perrett, D. I. (2013). Looking Like a Leader–Facial Shape Predicts Perceived Height and Leadership Ability. PLoS ONE 8(12): e80957. https://doi.org/10.1371/journal.pone.0080957

Renoult, J. P., Bovet, J., & Raymond, M. (2016). Beauty is in the efficient coding of the beholder. *Royal Society Open Science, 3*(3). https://doi.org/10.1098/rsos.160027

Rhodes, G. (2006). The evolutionary psychology of facial beauty. *Annual Review of Psychology, 57*, 199-226. https://doi.org/10.1146/annurev.psych.57.102904.190208

Russell, R., Porcheron, A., Sweda, J. R., Jones, A. L., Mauger, E., & Morizot, F. (2016). Facial contrast is a cue for perceiving health from the face. *Journal of Experimental Psychology: Human Perception and Performance, 42*(9), 1354-1362. https://doi.org/10.1037/xhp0000219

Said, C. P., & Todorov, A. (2011). A statistical model of facial attractiveness. *Psychological Science, 22*, 1183-1190. https://doi.org/10.1177/0956797611419169

Stephen, I. D., Law Smith, M. J., Stirrat, M. R., & Perrett, D. I. (2009). Facial Skin Coloration Affects Perceived Health of Human Faces. *International journal of primatology, 30*(6), 845-857. https://doi.org/10.1007/s10764-009-9380-z

Thornhill, R., & Gangestad, S. W. (1999). Facial attractiveness. *Trends in Cognitive Sciences, 3*, 452-460. https://doi.org/10.1016/S1364-6613(99)01403-5

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology*, 1-46. https://doi.org/10.1146/annurev-psych-113011-143831

Wolffhechel, K., Hahn, A. C., Jarmer, H., Fisher, C. I., Jones, B. C., & DeBruine, L. M. (2015). Testing the utility of a data-driven approach for assessing BMI from face images. *PLOS ONE, 10*(10), e0140347. https://doi.org/10.1371/journal.pone.0140347

Zebrowitz, L. A. & Rhodes, G. (2004). Sensitivity to “Bad Genes” and the Anomalous Face Overgeneralization Effect: Cue Validity, Cue Utilization, and Accuracy in Judging Intelligence and Health. *Journal of Nonverbal Behavior*, *28*, 167-185. https://doi.org/10.1023/B:JONB.0000039648.30935.1b